The molecular basis of nuclear genetic code change in ciliates

Catherine A. Lozupone, Robin D. Knight and Laura F. Landweber*

Background: The nuclear genetic code has changed in several lineages of ciliates. These changes, UAR to glutamine and UGA to cysteine, imply that eukaryotic release factor 1 (eRF1), the protein that recognizes stop codons and terminates translation, changes specificity. Here we test whether changes in eRF1 drive genetic code evolution.

Results: Database sequence analysis reveals numerous genetic code alterations in ciliates, including UGA → tryptophan in *Blepharisma americanum* and the distantly related *Colpoda*. We sequenced eRF1 from four ciliates: *B. americanum*, a heterotrich that independently derived the same eRF1 specificity as *Euplotes*, and three spirotrichs, *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax*, that independently derived the same genetic code as *Tetrahymena* (UAR → glutamine). Distantly related ciliates with similar codes show characteristic changes in eRF1. We used a sliding window analysis to test associations between changes in specific eRF1 residues and changes in the genetic code. The regions of eRF1 that display convergent substitutions are identical to those identified in a recently reported nonsense suppression mutant screen in yeast.

Conclusions: Genetic code change by stop codon reassignment is surprisingly frequent in ciliates, with UGA → tryptophan occurring twice independently. This is the first description of this code, previously found only in bacteria and mitochondria, in a eukaryotic nuclear genome. eRF1 has evolved strikingly convergently in lineages with variant genetic codes. The strong concordance with biochemical data indicates that our methodology may be generally useful for detecting molecular determinants of biochemical changes in evolution.

Address: Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544

Correspondence: Laura F. Landweber
E-mail: lfl@princeton.edu

Received: **20 October 2000**
Revised: **26 November 2000**
Accepted: **27 November 2000**

Published: **23 January 2001**

Current Biology 2001, 11:65–74

0960-9822/01/$ – see front matter

Background

The genetic code was once thought to be universal among all organisms: once fixed, any change – tantamount to rewiring a keyboard – would cause deleterious changes in every protein [1]. We now know that the code can change; alternative genetic codes are found in most mitochondrial genomes [2], the nuclear genomes of the eubacterium *Mycoplasma* [3], the yeast *Candida* [4], diplomonads [5], the green alga *Acetabularia*, and a variety of ciliates [e.g., 6–10]. Ciliates are remarkable in this respect. Several different code variants have arisen independently, even within a single class [11]. For example, *Tetrahymena* and *Paramecium*, class Oligohymenophorea, and *Oxytricha* and *Stylonychia*, class Spirotrichea, translate UAA and UAG as glutamine (using only UGA as stop) [6, 12]. *Euplotes*, also a spirotrich, instead translates UGA as cysteine, using UAA and UAG for termination [13]. *Blepharisma*, class Heterotrichea, uses UAA to encode stop [8]: before this study, the translation of UAG and UGA in this species was unknown. Although GenBank includes a separate translation table for *Blepharisma*, called the "Blepharisma code," in which UAA and UGA encode stop and UAG encodes glutamine [8], we find no support for the translation of UAG as glutamine or UGA as stop in this species. In a 1995 study, Baroin Tourancheau et al. [11] sequenced the α-tubulin and phosphoglycerate kinase genes for members of 6 of the 10 currently recognized ciliate classes and suggested that members of the class Litostomatea and the heterotrich *Stentor coeruleus* may use the standard genetic code, whereas the karyorelictid *Loxodes striatus*, the heterotrich *Condylostoma magnum*, and the nassophorean *Zosterographus sp.* appear to use UAA and UAG to encode glutamine. These data indicate that alteration of the ciliate genetic code was not a single, ancient event, as initially supposed [14], but a relatively common event in ciliates [11].

Changes in the genetic code involve changes in tRNAs, tRNA-modifying enzymes, or release factors. These crucial components of the translation apparatus have changed in ciliates, conferring new meanings to specific codons. Altered tRNAs in organisms with variant genetic codes have provided some insight into the mechanism of genetic code changes. In addition to normal $tRNA^{Gln}$, *Tetrahymena thermophila* has two unusual glutamine-specific $tRNA^{Gln}$ with anticodons complementary to UAA and UAG: these unusual tRNAs arose by duplication and divergence of the canonical $tRNA^{Gln}$ [7]. *Euplotes octocarinatus* has only

one $tRNA^{Cys}$ that translates UGA efficiently, despite G:A mispairing at the first anticodon position [15]. This change requires loss of release factor specificity for UGA and/or high concentrations of $tRNA^{Cys}$ relative to other tRNAs.

Because all known ciliate code changes alter stop codon meanings, the eukaryotic release factor 1 (eRF1) must have evolved alternate specificities. In eukaryotes, eRF1 recognizes the three standard stop codons in mRNA at the ribosomal A site and terminates translation by peptidyl tRNA hydrolysis. Archaea have an eRF1 homolog, aRF1, which is highly conserved across domains, and aRF1 even functions with eukaryotic ribosomes [16].

Although eRF1 sequences from organisms with altered termination are of particular interest, only one eRF1 sequence was in GenBank for an organism with a nonstandard genetic code (the ciliate *Tetrahymena thermophila* [17]). However, this eRF1 sequence suggested a mechanism for the specificity change: the NIKS motif, conserved across all eukaryotes, is NIKD in *Tetrahymena*. Together with the recent crystal structure of human eRF1 [18] and mutational evidence that changes adjacent to NIKS abolish stop codon recognition in vitro [19], we suggested that this specific mutation might be the molecular cause of *Tetrahymena's* altered genetic code [20].

In order to test this hypothesis and to further examine the biochemical basis for altered stop codon recognition in ciliates, we sequenced the complete gene encoding eRF1 in three spirotrichs, *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax*, that independently derived the same genetic code as *Tetrahymena*. For comparison, we also sequenced the gene in *Blepharisma americanum*, an early diverging ciliate that uses UAA as stop [8], and found evidence that this species uses UGA to encode tryptophan. Using a novel statistical approach that uses sliding window analysis to associate changes in specific regions of the protein with changes in the genetic code, as well as mapping some of the convergent amino acid substitutions in lineages that independently evolved the same eRF1 specificity onto the protein crystal structure, we identify several candidate amino acid residues or regions of the protein that may underlie the altered codon specificity of eRF1 and genetic code change in ciliates.

Results

Isolation of eRF1

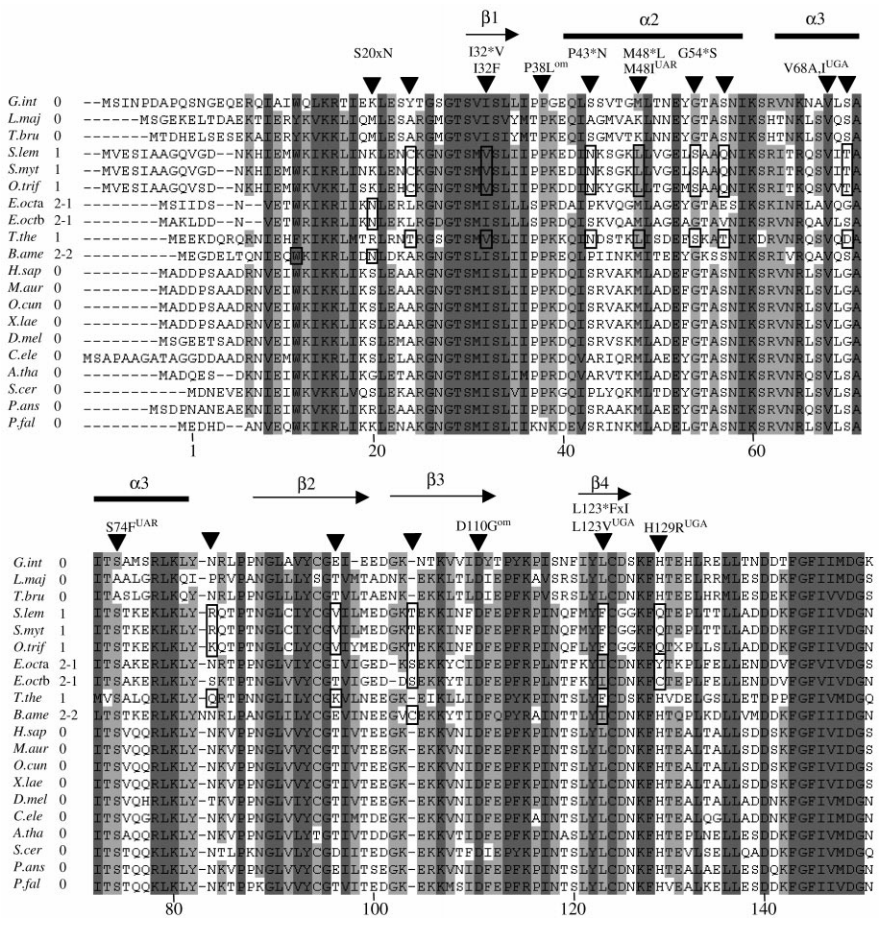
We determined the complete macronuclear sequence of *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax* eRF1 gene-sized chromosomes, which are all predicted to encode proteins 445 amino acids long, as well as *Blepharisma americanum* eRF1, predicted to encode a 436 amino acid protein. *S. lemnae*, *S. mytilus*, and *O. trifallax* each appear to have a phase I intron, 32, 32, and 38 nucleotides long, respectively, at position 78 in the amino acid alignment (Figure 1); *B. americanum* has no intron at this position. The eRF1 gene has 2 in-frame UGA codons in *B. americanum* and numerous in-frame UAR codons in *S. lemnae*, *S. mytilus* and *O. trifallax*.

Genetic database analysis

Protein sequence data were available in the database for 7 of the 10 currently recognized [21] classes of ciliates and for a member of the order Amorphoridae classified as *sedis mutabilis* in the subphylum Intramacronucleata [21]. We analyzed, for the first time, data for members of the class Colpodea and for *Nyctotherus*, of the order Amorphoridea. In addition, we expanded analysis of the six previously studied classes [11] to include more species allowing better definition within these groups. Figure 2 is a composite tree assembled from the literature, using both 28S large subunit [11, 22] and 18S small subunit [23, 24, 25] congruent rDNA phylogenies, for the purpose of character mapping of genetic codes. It lists all of the genera for which information on genetic code usage is available, except for many of the Spirotrichs analyzed, which form a monophyletic group with *Stylonychia*, *Oxytricha*, and *Urostyla*, and all appear to use the same code. These data support that the classes Oligohymenophorea, Spirotrichea, and Litostomatea are each monophyletic. The heterotrichs and the karyorelictids form an early branching monophyletic group, but the relationship of the rest of the classes is largely unresolved. There is some support, however, that the classes Nassophorea, Colpodea, and Oligohymenophorea form a monophyletic group (Figure 2). Members of the classes Karyorelictea and Nassophorea are grouped together by morphological data only [26, 27] and are placed on the tree based on the rRNA sequences of *Zosterograptus* and *Loxodes*, respectively. Table 1 summarizes the database analysis and supplemental Table 1 provides more detail (see Supplementary material).

We found evidence in two distantly related lineages for UGA reassignment to tryptophan (Figure 1 and supplemental Figure 1). The gene for mitotic cyclin-like protein in *Colpoda inflata* contains an in-frame UGA in the position of a conserved tryptophan (supplemental Figure 1). Analysis of five partial protein sequences revealed no in-frame UAR codons in this class (Table 1). In the heterotrichs, members of the genera *Stentor* and *Eufolliculina* both had multiple in-frame UGA codons, and *Eufolliculina* and *Blepharisma* use UAA to encode stop (Table 1; [8]). Alignment of the *Stentor* and *Eufolliculina* proteins containing in-frame UGA codons with orthologs retrieved using BLAST did not suggest which amino acid UGA was coding for because the alignment was in variable regions of the proteins. However, the *B. americanum* eRF1 sequence generated in this study had two in-frame UGA codons, both in the location of conserved tryptophan residues (Figure 1 and supplemental Figure 1). Based on the close

Figure 1



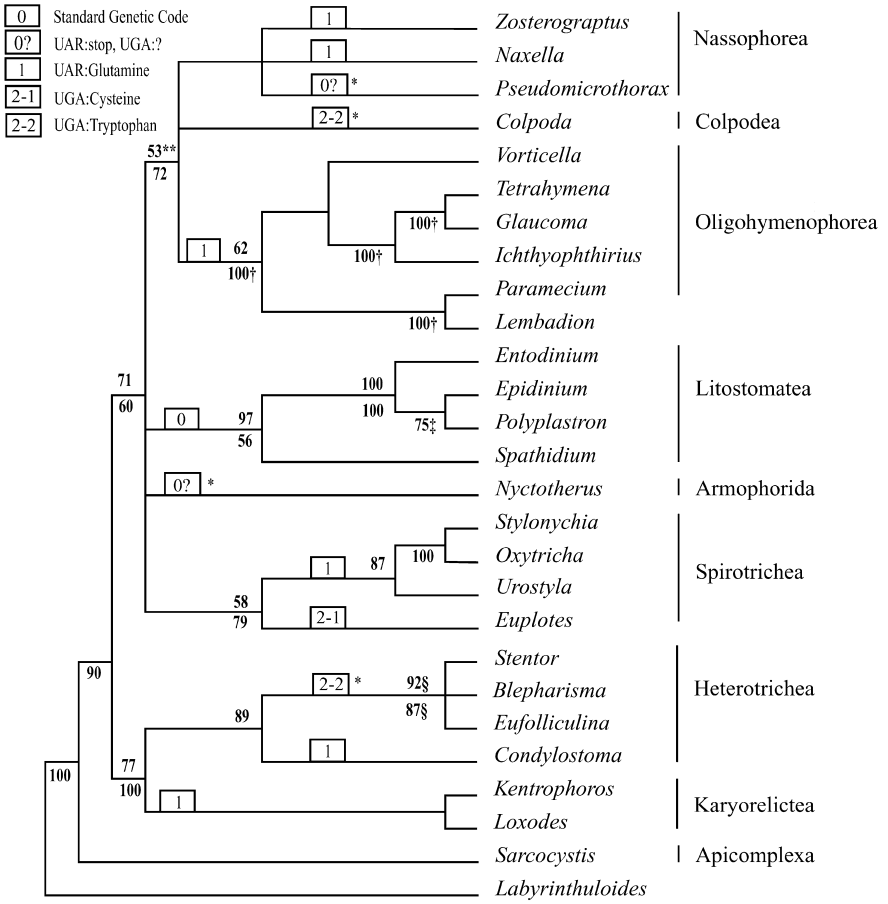
Alignment of the N terminus of all available eRF1 proteins. Areas of the alignment corresponding to structural features of domain 1 (α helices and β sheets) are indicated. Genetic code usage of each species is listed next to the abbreviated Latin name using the code notation described in Figure 2. Residues marked with an arrow are shown in Figure 3 and those that Bertram et al. [30] identified are indicated with their notation plus om (omnipotent suppression), UAR, or UGA to indicate the new recognition suppression of the mutant. Sites of interest are boxed and sites of convergent evolution between *Stylonychia/Oxytricha* and *Tetrahymena* (*) or *Euplotes* and *Blepharisma* (x) are annotated with the symbols (* or x) between the residues that mutated and the amino acid position in yeast (e.g., L123*F convergently changed to F at yeast position 123 in *Stylonychia/Oxytricha* and *Tetrahymena*). The boxed W at position 11 in the *Blepharisma* sequence is encoded as UGA (Supplemental Figure 1). Residues shaded in dark gray are highly conserved, variable residues are unshaded, and intermediate residues are light gray. Numbering according to the yeast sequence, *Saccharomyces cerevisiae* (*S.cer*, CAA51935) as in [30]. *G.int.*, *Giardia intestinalis* (AF198107 [41]); *L.maj*, *Leishmania major* (CAB77686); *T.bru*, *Trypanosoma brucei* (AAF86346); *S.lem*, *Stylonychia lemnae* (AF31784, this study); *S.myt*, *Stylonychia mytilus* (AF31783, this study); *O.tri*, *Oxytricha trifallax* (AF31782, this study); *E.octa*/b, *Euplotes octocarinatus* eRF1a and b [31]; *T.the*, *Tetrahymena thermophila* (P46055); *B.ame*, *Blepharisma americanum* (AF31781, this study); *H.sap*, *Homo sapiens* (P46055); *M.aur*, *Mesocricetus auratus* (X81626); *O.cun*, *Oryctolagus cuniculus* (AB029089); *X.lae*, *Xenopus laevis* (P35615); *D.mel*, *Drosophila melanogaster* (AAF51574); *C.ele*, *Caenorhabditis elegans* (T31907); *A.tha*, *Arabidopsis thaliana* (P35614); *P.ans*, *Podospora anserina* (AAC08410); and *P.fal*, *Plasmodium falciparum* (AAC71899).

phylogenetic relationship of *E. uhligi* and *S. coeruleus* to *B. americanum* [11, 28] and the presence of in-frame UGA codons in protein-coding genes for these species, it is likely that *E. uhligi* and *S. coeruleus* use UGA to encode tryptophan as well. This genetic code has been observed in mitochondrial genomes [2] and in the eubacterial *Mycoplasma* [3] but has never before been described for eukaryotic nuclear genomes. The heterotrichs are also an example of members of the same class using two different divergent codes: the early diverging heterotrich *Condylostoma magnum* has in-frame UAA and UAG codons at a position where conserved glutamines usually occur in orthologs [11].

Our results support the finding of Baroin Tourancheau et al. [11] that the reassignment of UAR to glutamine appears to have occurred multiple times independently in ciliates. Within the classes Oligohymenophorea and Spirotrichea, *Tetrahymena*, *Paramecium*, *Stylonychia*, *Oxytricha*, and *Glaucoma* all use this code [11]. We find evidence for use of UAR as glutamine in other oligohymenophorans, *Ichthyophthirius multifilis* and *Vorticella convallaria*, as well as many spirotrichs (Table 1). This shows that use of this code is common to members of these classes, except for the spirotrich *Euplotes*, which uses UGA to encode cysteine and UAR as stop [13]. The three other lineages that use UAR to encode glutamine are *Loxodes striatus* (Karyorelictea), *Zosterographus* (Nassophorea), and *Condylostoma* (Heterotrichea) [11]. We strengthen these findings by showing in-frame UAR codons in the karyorelictid *Kentrophoros sp.* and the nassophorean *Naxella sp.* (Table 1).

The class Nassophorea also displays use of different genetic codes within a ciliate class, although no molecular data are available on the monophyly of this class. While

Figure 2



Mapping of genetic code changes (boxes) mapped onto a composite phylogenetic tree of ciliate genera compiled from congruent published analyses of 28S ([11] bootstrap values indicated above the line in boldface; [22] marked **) and 18S ([23] bootstrap values below the line in bold; [25] marked †; [24] ‡) rRNA sequences. Bootstrap values marked § represent 28S and 18S studies, respectively, that grouped *Blepharisma* with *Stentor* 92% [11] and with *Eufolliculina* 87% [28]. A single asterisk indicates lineages with a code change described for the first time in this study.

in-frame UAR codons are present in *Zosterograptus sp*. and *Naxella sp*., three protein sequences in *Pseudomicrothorax dubius* have no in-frame stop codons and use UAG and UAA to terminate translation. The coding of UGA in this species was ambiguous. The sequence of the hydrogenase protein in *Nyctotherus ovalis*, of the order Amorphorida, has no in-frame stop codons and uses UAA to encode stop. Additional DNA sequences from these two species should be examined for evidence of UGA reassignment to an amino acid.

The class Litostomatea is the only group of ciliates for which strong evidence suggests use of the standard genetic code. None of the 55 protein sequences available for litostomes had in-frame stop codons, and translation was terminated with either UAA or UGA (Table 1 and supplemental Table 1).

Identification of convergent changes

The changes in ciliate genetic codes involve two types of altered stop codon recognition: recoding of either UGA or both UAA and UAG. For each of these two types of change, we have eRF1 sequences from representatives of two different lineages in which the change has occurred: the reassignment of UGA in *Euplotes* and *Blepharisma* and UAR in *Stylonychia/Oxytricha* and *Tetrahymena*. The phylogenetic analysis indicates that both of these changes probably occurred independently. For the UGA change, it is strongly supported that *Euplotes* is more closely related to many genera that use UAR to encode glutamine, such as *Stylonychia*, *Oxytricha*, and *Tetrahymena*, than it is to *Blepharisma* (Figure 2). *Blepharisma*, likewise, is closely related to *Condylostoma magnum*, which also uses UAR to encode glutamine [11]. There is weaker evidence that the change of UAR to glutamine occurred independently in the spirotrichs that use this code and the oligohymenophorans: *Euplotes*, which uses UGA to encode cysteine, appears to have diverged early from a monophyletic lineage that includes the other spirotrichs, and *Colpoda*, in which we find evidence of UGA use to encode tryptophan, appears to form a monophyletic group with the oligohymenophorans (Figure 2). Both of these relationships are

Table 1

Organism	#	3′	aa	UAG	UAA	UGA	stop	Code
Nassophorea	6							0?,1
Naxella	1	0	408	3	2	0	(UGA)	1
Zosterograptus	1	0	382	0	6	0		1
Pseudomicrothorax	3	3	1770	0	0	0	UAA,UAG	0?
Colpodea	4							2-2
Colpoda	4	0	1253	0	0	1		2-2
Oligohymenophorea	6+							1
Vorticella	1	1	181	1	0	0	UGA	1
Glaucoma	2	0	449	3	1	0	(UGA)	1
Ichthyophthirius	2	1	517	1	23	0	UGA	1
Lembadion	1	1	351	0	6	0	UGA	1
Litostomatea	55							0
Entodinium	47	45	9447	0	0	0	UAA,UGA	0
Epidinium	2	1	737	0	0	0	UAA	0
Polyplastron	3	3	930	0	0	0	UAA	0
Spathidium	3	0	1194	0	0	0		0
Armophorida	2							0?
Nyctotherus	2	1	1742	0	0	0	UAA	0?
Spirotrichea	14+							1, 2-1
Pleurotricha	1	0	798	5	31	0	(UGA)	1
Paraurostyla	1	0	966	10	24	0	(UGA)	1
Uroleptus	1	0	973	10	26	0	(UGA)	1
Urostyla	4	2	2606	15	33	0	UGA	1
Histriculus	2	1	755	0	1	0	UGA	1
Halteria	3	1	1697	7	5	0	UGA	1
Holosticha	1	0	966	11	19	0	(UGA)	1
Hypotrichida	1	0	373	0	0	0		
Heterotrichea	32							1, 2-2
Blepharisma	10	2	689	0	0	2	UAA	2-2
Eufolliculina	12	12	3250	0	0	4	UAA	2-2?
Stentor	8	0	1841	0	0	3		2-2?
Condylostoma	1	0	380	1	4	0	(UGA)	1
Karyorelictea	2							1
Kentrophoros	1	0	408	1	0	0		1
Loxodes	1	0	380	6	0	0		1

A summary of ciliate class/genera, total number of protein sequences available; number of sequences that were complete at the 3′ end allowing determination of stop codon usage; total number of amino acids (aa) analyzed; number of in-frame UAA, UAG, and UGA codons, the codon(s) used to terminate translation where available (codons in parentheses are inferred from the data); and the genetic code used: (0) standard genetic code, (0?) UAA and UAG used as stop and translation of UGA is unknown (1) UAR:glutamine, (2) UGA:amino acid, (2-1) UGA:cysteine, (2-2) UGA:tryptophan.

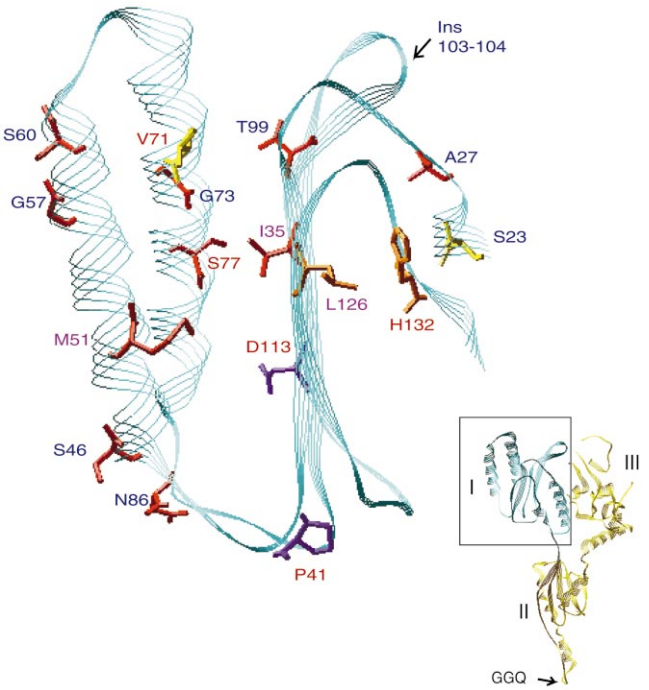
supported in numerous 18S and 28S rRNA trees, generally with greater bootstrap support for studies based on complete 18S small subunit (e.g., [23, 24, 28]) than partial 28S large subunit [11, 22] sequences. Morphological data also support the classification of *Euplotes* as a spirotrich [29]. In addition, phylogenetic analysis of α-tubulin gene sequences weakly supports both of these relationships, and an analysis of phosphoglycerate kinase gene sequences strongly supports the branching of *Euplotes* with *Oxytricha* (97% bootstrap support), though data for *Colpoda* were unavailable [22].

If these phylogenetic inferences are correct, one could still argue that an ancient change of UAR to glutamine in a common ancestor of the spirotrichs and oligohymenophorans might have been followed by UGA reassignment in *Euplotes* and *Colpoda*. It is much more likely, however, that the transition to these alternate genetic codes proceeded from a standard code background. The evolution of the *Euplotes* code from an ancestor that used UAR to encode glutamine, for instance, would potentially require loss of extra glutamine tRNAs, eRF1 mutations that decrease UAR readthrough and increase UGA readthrough, and mutations enabling a $tRNA^{Cys}$ to read UGA. For these reasons, we propose that the UAR-to-glutamine change most likely emerged independently in the spirotrich and oligohymenophoran lineages. We next ask whether there are convergent changes that could confer the altered specificity for both UAR and UGA reassignment. In other words, are there states of particular sites in eRF1 such that all and only ciliates with the same altered translation termination share those particular states? This analysis is itself independent of the direction of change inferred in Figure 2.

We wrote a C program (available from the authors) to scan a set of sequences labeled according to arbitrary criteria (in this case, the type of stop codons used by the species) and to partition amino acid identities at each site of an aligned set of sequences according to which labels

have which sites. We find that the following states are unique to particular translation states (numbered according to yeast eRF1; amino acids to the left of a number occur in organisms with the standard code and the amino acids to the right of the number occur in organisms with altered stop codon specificity): UAR = Stop: (KRGMS)20N, (.)103–104(CS), **L123I**, (NGT)213(DES), N262(ST), (AQ)279(ER), (DEIMQ)366(GV); UGA = Stop: (AY)24(CT), **I32V**, (PSA)43N, **(MK)48L**, G54S, (ES)57(QT), (GS)70(DT), (NPT)83KQR, (ETD)96(KV), (.)103–104T, **L123F**, (SAGKQ)273T, (NDG)322E, (.NDGQ)332(.K), (DECQS)343(NT). Remarkably, 14 of these 22 convergent changes map to domain 1, the putative codon recognition domain [18], and 4 of these changes in 3 different positions (in boldface) are the same changes recently identified by mutational screens for enhanced termination suppression in yeast [30]! Our alignment spanned 417 residues: the probability that we would hit 3 or more of the same residues that the mutational study identified by chance, if the changes were independent, is 0.014. (The mutational study identified 10 of 417 residues as being functionally important. Of these, 9 were in the region covered by our alignment. Since we identified 20 sites as having states unique to particular genetic codes, the 2 × 2 contingency table has counts of 3, 6, 17, and 319 for identification in both studies, in [30], our study, and neither study respectively. The probability that these results are independent is very low [G test for independence applying the Williams correction: df = 1, G = 4.82, $P_{1\text{-tailed}} = 0.014$]. Thus, our identification of several of the same sites is unlikely to be a coincidence.)

Figure 3



Human eRF1 structure of domain 1. The entire protein ribbon structure of human eRF1 (inset) is shown with domain 1 boxed in blue and domains 2 and 3 in yellow (PDB accession code 1DT9 [18]). Highlighted residues in domain 1 have been identified in this study (blue text), mutational analysis in yeast (red text [30]), or both (purple text). Numbers correspond to human eRF1, +3 from yeast eRF1 and Figure 1 (e.g., human M51 is homologous to yeast M48). Mutations in red residues are associated with UAR suppression; yellow, UGA suppression; and orange, both.

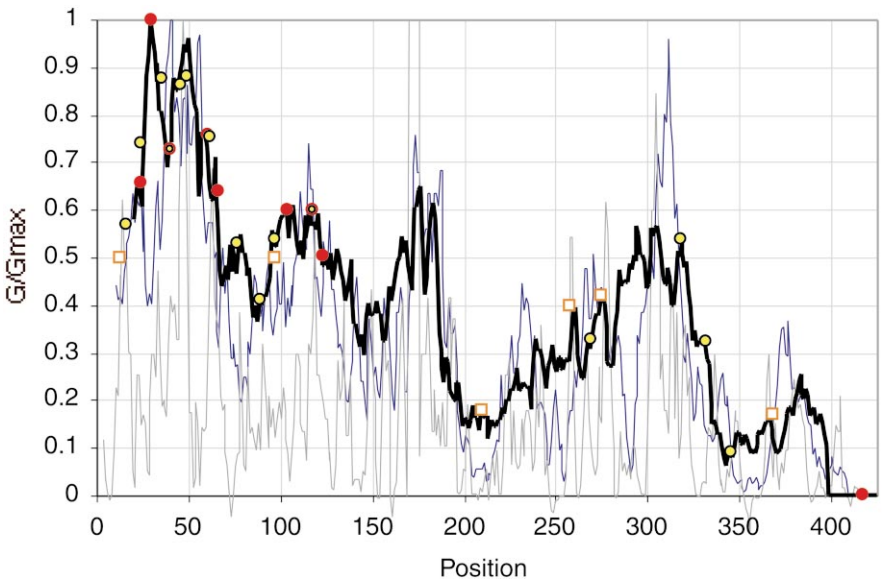
Bertram et al. [30] identified ten changes in domain 1: three changes in mutants exhibiting UGA suppression (V68I, L123V, and H129R) increased UGA readthrough, and three changes in mutants with UAG suppression (M48I, S74F, D110G) increased UAR readthrough. I32F was a potential UGA suppressor and the remaining three isolates were omnipotent suppressors. Our evolutionary approach agrees surprisingly well with these mutational data: the UGA suppressors *Euplotes* and *Blepharisma* have a convergent change from L to I in the homologous position of L123V, and *Euplotes* [31] has a Y or C at the position of H129R, a conserved histidine in all taxa except ciliates with alternate codes (Figure 1). V68I is at a conserved valine across all ciliates, indicating that this site either has not been explored by evolution or is fixed by other constraints. L123V is also a site of a convergent change from L to F in *Stylonychia/Oxytricha* and *Tetrahymena* and is conserved in all other taxa; *Stylonychia* and *Oxytricha* have a nonconservative change from H to Q at site H129R. The homologous residues L126 and H132 in human eRF1 flank a hydrophobic pocket that may recognize the stop codon second position in the model in [30] (Figure 3). As this position distinguishes UAR from UGA, we propose that these changes, in particular the convergent changes at position 123 (Figure 1), play a critical role in the evolution of release factor specificity in ciliates. There is only one additional convergent change in domain 1 of *Euplotes* and *Blepharisma*. In contrast, we identified ten sites in domain 1 that changed in both *Stylonychia/Oxytricha* and *Tetrahymena*. Of these, seven are in α helices 2 and 3 (Figures 1 and 3) and five mutate to the same residue. Two convergent changes, I to V and M to L, are at the same site as I32F and M48I UAR suppression mutations [30], and site 32 is conserved across all taxa but the ones using this code. An important next step would be to engineer these changes into yeast eRF1 and test whether the resulting protein could rescue a *Tetrahymena* eRF1 knockout.

Sliding window analysis: associating eRF1 changes and termination changes

The analysis of individual residues is sometimes difficult to interpret, because apparent associations between change

Figure 4

Sliding window analysis of eRF1 and genetic code change. This figure shows the robustness of peaks to changes in window size, showing windows of size 5 (gray line), 20 (blue line), and 40 (black line). Peaks indicate regions in which changes are maximally associated with genetic code change. The x axis is position in our alignment (Figure 1); the y axis is relative G value (scaled relative to maximum G value for that window size). The first two peaks lie in domain 1, the third in domain 2 near the GGQ motif, and the fourth in domain 3, which binds eRF3 [18]. Red dots mark mutations in yeast eRF1 that affect stop codon readthrough (marked on the 40 residue line); UAR→Gln, yellow dots; UGA ≠ stop, orange boxes. An interactive Excel file that allows the user to view the results with different window sizes is available at ftp://rnaworld.princeton.edu/pub/export/windowresults2.xls



at a single site and change in the genetic code could occur by chance. By pooling counts for a contiguous stretch of sequence, it is possible to test for association between changes in that region and changes in the genetic code. This increases confidence that particular regions are important, at the expense of making inferences about individual sites.

Figure 4 shows results for three window sizes (5, 20, and 40 residues) along with mutations found to affect decoding in yeast eRF1 [30] and those identified in this study. There are four peaks that are stable across a broad range of window sizes. The first two of these peaks correspond to domain 1, and the third one corresponds to the region immediately before the universal GGQ motif. Additionally, the decoding mutants in yeast and the convergent changes identified in this study cluster around the first two peaks.

Discussion

Our database analysis extends both the number of independent code changes in ciliates and the types of code change reported for eukaryotes. Both UAR and UGA have been reassigned independently many times in different lineages of ciliates. Whether ciliates are particularly prone to changes in the genetic code or if they are they just more diverse than other microbial eukaryotes remains open. UGA has been reassigned to both cysteine and tryptophan, showing that either of two tRNAs can expand its decoding ability to cover this codon. The UGA → tryptophan change also occurs in bacteria and mitochondria. Is UGA particularly prone to reassignment?

We observed many fewer convergent eRF1 substitutions in *Euplotes* and *Blepharisma*, which suggests either that independent substitutions may assist stop codon reassignment in these lineages or, more intriguingly, that UGA suppression may be easier to effect than UAR suppression, a conclusion that is also supported by the fact that Bertram et al.'s initial screen for unipotent suppressor phenotypes only identified UGA suppressors [30]. The isolation of UAR suppressors required a plasmid containing a mutant $tRNA^{Ser}$ with weak UAR nonsense suppressor activity.

We initially set out to test the hypothesis that a change of S to D in *Tetrahymena*, within the otherwise universally conserved NIKS motif of domain 1, was responsible for altering the specificity of eRF1 in organisms with this code. We found that *Stylonychia* and *Oxytricha*, which also translate UAR as glutamine, do not have any deviations in the NIKS motif; nor does *Blepharisma*. One of the two release factor genes in *Euplotes octocarinatus*, however, has the sequence SIKS in this region [31]. This indicates that changes in this motif are not necessary for altered stop codon recognition, as we initially proposed, although they may be sufficient. However, several different changes adjacent to this region are heavily implicated in variant codes. Construction of specific mutants in yeast eRF1 will allow finer mapping of the critical residues and perhaps indicate how many ways eRF1 can mutate to generate the same changes in the genetic code.

Changes in ciliate termination may depend on less than 6 residue changes (for loss of UGA recognition) or less than 15 residue changes (for loss of UAR recognition) in eRF1 on the basis of convergent mutations in lineages that evolved the same changes independently. Despite

potentially requiring more mutations in eRF1, UAR has changed to glutamine in at least two other completely independent groups (diplomonads and algae [5, 32]); eRF1 sequences from these species may reveal whether this change in the genetic code has a unique convergent molecular basis.

Most changes in the genetic code involve termination: this may be because stop codons are rare, occurring only once per gene, and so changes in termination are likely to be less deleterious than change in sense codons. This would be particularly true for those species of ciliates whose genes reside on gene-sized chromosomes and/or have short 3′ untranslated regions. In addition, termination is a competition for stop-codon-containing ribosomal A sites between release factors and tRNAs. Consequently, relatively small changes either in the tRNAs or in eRF1 may shift this balance toward partial or complete readthrough in some cases. For instance, *Bacillus subtilis* uses in-frame UGA codons extensively to encode tryptophan; however, this readthrough is inefficient, and UGA is also used as a stop codon [33, 34]. The abundance of stop codon reassignments relative to amino acid codon reassignment, however, could also be an observer bias. In-frame stop codons are much easier to detect in protein coding sequences than amino acid replacements, especially if the latter have similar properties.

Is there any pattern to the identities of the amino acids to which the stop codons are reassigned? The reassignment of UAR to glutamine can be explained either by a transition mutation at the third anticodon position of $tRNA^{Gln}$ (which normally recognizes CAA and/or CAG), alteration in the tRNA elsewhere to enhance G-U mispairing at the first codon position, or both. The second mechanism implies a period of ambiguous translation; interestingly, *Tetrahymena* $tRNA^{Gln}$ contain specific changes distant from the anticodon that increase G-U wobble when introduced into the equivalent tRNA in yeast [35]. Similarly, reassignment of UGA to cysteine and tryptophan can be explained in terms of expansion of wobble in existing tRNAs [35].

We identified convergent changes in three of the exact same sites in eRF1 independently identified by mutational screens in yeast [30]. Additionally, the yeast mutants cluster heavily around the first two major peaks we identified in domain 1 by sliding window analysis. This approach also identified a sharp peak at about position 175, immediately adjacent to the GGQ motif. This motif mimics the tRNA CCA acceptor stem and probably interacts with the peptidyl transferase center of the ribosome to release the nascent peptide [18]. In a previous study, Karamyshev et al. [17] found that *Tetrahymena* eRF1 does not work with yeast ribosomes: this was surprising, since even the distantly related aRF1 from the archaeon *Methanococcus jannaschii* is active with eukaryotic ribosomes [16], and suggests that these structures in *Tetrahymena* have coevolved.

The sliding window analysis proves surprisingly powerful in identifying the sites already known to be particularly important in eRF1 function: domain 1 and the GGQ motif. This type of analysis should be applicable to any character that has changed multiple times independently in different taxa: possible examples include changes in signaling pathways, recruitment of paralogs to new functions, and changes in pathogenicity or host range. Identification of molecular correlates of evolutionary change in this manner may greatly assist corresponding biochemical analyses.

These results show the power of comparative evolutionary techniques in identifying important sites in functional molecules, as Woese et al. showed decades earlier with rRNA [36]. Finding several species that differ naturally in some respect, and determining molecular correlates of this change, may yield the same results as screening for mutants in a single species. This approach may even be more powerful, as change-of-function mutants are often lethal and may not be picked up efficiently in a selection. We predict that many of the changes that we identify above, when introduced into yeast eRF1, will be lethal because of efficient stop codon readthrough.

Conclusions

We have shown that genetic code change in ciliates is even more pervasive than previously thought, with multiple independent lineages evolving changes in stop codon recognition requiring changes in release factor eRF1. We report the first use of UGA as tryptophan in a eukaryotic nuclear genome, in two distinct lineages; therefore, even this change has occurred more than once in ciliates. We detect striking instances of convergent evolution in the amino acid sequence of eRF1, with distantly related lineages with variant codes displaying characteristic substitutions in eRF1, implying that some of these substitutions may drive evolution of the genetic code. The strong agreement between our results and nonsense suppression screens in yeast [30] strengthens our conclusions, as well as the robustness of the statistical analysis. Thus, the combination of unique residue identification and sliding window analysis provides a powerful approach to detecting the molecular determinants of convergent evolutionary change.

Materials and methods

Release factor sequencing and analysis

Macronuclear DNA from *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax* were a gift from David Prescott (University of Colorado, Boulder). DNA was extracted from *Blepharisma americanum* (Culture Collection of Algae and Protozoa, CCAP 1607/1) using a DNeasy Tissue Kit (QIAGEN). We then aligned all eRF1 sequences available from GenBank and designed degenerate primers flanking the codon recognition domain of eRF1. The forward primer is eRF1100F, ATG(AG)T(TA)(TA)C

(ATC)TT(GA)(GA)T(TC)AT(TC)CC(TA)CC and the reverse primer is eRF1799R, AT(KT)GC(TC)TGKTTKAA(AT)CCKTTNTC(AT)CC(AT)CC (AT)CCKTA. K is a nonstandard base that binds both C and T (Glen Research). We amplified the region encoding the eRF1 protein fragment using 40 cycles of PCR (94°C, 25 s; 50°C, 20 s; 72°C, 1 min; final extension , 72°C, 10 min). We then amplified the 5′ and 3′ ends of the macronuclear eRF1 gene for *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax* using gene specific primers SORF415F [TATTTTTGC GGTGGTAA(AG)TTCCAGACTGA], SORF680R [(AT)AGTCTCTTAT CGAGCAT(GA)TCAGTCTCA], SORF581F [ACAGAAAGGGAGGT CA(AG)TC(AT)TCAGTCAG], and SORF604R [CTGA(TA)GA(CT)TG ACCTCCCTTTCTGTGCTT] and telomere-specific primers in telomere suppression PCR as described elsewhere [37].

We recovered the 5′ and 3′ ends of the *Blepharisma americanum* eRF1 cDNA using anchor PCR primers and protocols as described in [38] with total RNA extracted using TRIzol reagent (GIBCO–BRL), Superscript II reverse transcriptase (GIBCO–BRL), and primers specific for *B. americanum* eRF1: BGSP1F (TGCCAGGCTGCGTATGGAGTCC), BGSP 2F (CATTGCTGGGTCAGCTGAGTTC), BGSPRTR (AAATCAGACC GTTTGCTGGGAG), BGSP1R (CGCTCTTTTGTGGAGGTAAGAG), and BGSP2R (ACAGCCTGCCTGACGATTCTTG).

All PCR products were cloned using a TOPO TA Cloning Kit (Invitrogen) and multiple clones for each fragment were sequenced at the Syn/Seq Facility of Princeton University. The complete *B. americanum* sequence was confirmed by PCR amplification and sequencing from total DNA. Sequences have been deposited in GenBank (AF31781–AF31784).

We aligned sequences with all known eRF1 homologs using Pileup in Seqlab (Wisconsin GCG version 10.1) and identified nonconservative mutations in otherwise conserved regions of the codon recognition domain in species with nonstandard genetic codes. Mutations common to all and only ciliates with one of the two specific types of change in termination (using either UGA only or UAR only) were of particular interest here.

Additionally, we tested whether changes in specific areas of the eRF1 protein were unusually associated with changes in the genetic code. This relied on the phylogeny, with its inferred ancestral states. We constructed a maximum parsimony phylogeny using the Paupsearch function in Seqlab (Wisconsin GCG version 10.1) and a heuristic tree search using the default settings. We excluded from this analysis the 5′ and 3′ ends of the sequences, as well as 2 residues for which the amino acid identity of *Blepharisma americanum* was ambiguous and a 13 residue region in which a 9 residue insert in one of the sequences left the alignment unresolved. We consistently observed the artificial grouping of *Stylonychia* and *Oxytricha* (class Spirotrichea) with *Tetrahymena* (class Oligohymenophorea) instead of *Euplotes* (class Spirotrichea), but exclusion of five amino acid positions in domain 1, in which there was convergent evolution between *Stylonychia*/*Oxytricha*, and *Tetrahymena*, restored the monophyly of the spirotrichs. These residues were added back to the alignment after performing phylogenetic analysis, and the ancestral states of these residues were inferred based on parsimony [39].

For every edge linking two nodes in the phylogeny, we inferred whether a change in the code had taken place and, for each position in the alignment, whether the amino acid at that position had changed. This gave a table of four counts for each position, according to whether or not changes (in code and sequence) had occurred. It was thus possible to test for independence between code changes and changes at each position in eRF1; if change at a particular position caused changes in the code, the two variables would not be independent. We used the G test for independence [40].

Because there are relatively few sequences, individual residues typically did not give significant associations. In order to identify regions of sequence associated with genetic code change, we performed sliding window analysis, combining counts from contiguous regions of size n ($1 \leq n \leq 100$) to identify regions in which changes were consistently implicated with changes in the genetic code. This analysis was performed using custom programs written in C and Microsoft Excel–hosted VBA.

Genetic database analysis

We analyzed all ciliate protein coding sequences available in genetic databases, using the Seqlab editor of the GCG Wisconsin Package (version 10.1). We excluded sequences from *Tetrahymena*, *Paramecium*, *Stylonychia*, *Oxytricha*, and *Euplotes*, since the genetic code of these genera has been well characterized. We translated sequence data using the standard genetic code and recorded the number of in-frame stop codons and the type of codon actually used for terminating translation in each gene. Wherever possible, we aligned protein sequences displaying in-frame stop codons with orthologs from related species, allowing us to infer the new meaning of the former stop codon. Using congruent published molecular phylogenies based on 18S small subunit [23–25, 28] and 28S large subunit rRNA [11, 22] sequences (Figure 2), we then made a composite phylogenetic tree of the ciliates with known genetic code usage. Molecular data were available to assess the relationships of all groups except the nassophoreans and the karyorelictids, which were grouped according to morphological characters [26, 27].

Supplementary material

Supplementary material including partial eRF1 and cyclin sequence alignments showing evidence of UGA usage at conserved tryptophans and a table recording in-frame stop codon occurrence for the data summarized in Table 1, is available at http://current-biology.com/supmat/supmatin.htm.

Acknowledgements

We gratefully acknowledge David Prescott for the gift of *Stylonychia* and *Oxytricha* DNA and Klaus Heckmann for sharing the *Euplotes octocarinatus* sequences before publication. We also thank Stephen Freeland, Christina Burch, and David Ardell for discussion.

References

1. Crick FH: **The origin of the genetic code.** *J Mol Biol*1968, **38:**367-379.
2. Barrell BG, Bankier AT, Drouin J: **A different genetic code in human mitochondria**. *Nature* 1979, **282:**189-194.
3. Yamao F, Muto A, Kawauchi Y, Iwami M, Iwagami S, Azumi Y, *et al.*: **UGA is read as tryptophan in *Mycoplasma capricolum*.** *Proc Natl Acad Sci USA* 1985, **82:**2306-2309.
4. Santos MAS, Tuite MF: **The CUG codon is decoded in vivo as serine and not leucine in *Candida albicans*.** *Nucleic Acids Res* 1995, **23:**1481-1486.
5. Keeling PJ, Doolittle WF: **A non-canonical genetic code in an early diverging eukaryotic lineage.** *EMBO J* 1996, **15:**2285-2290.
6. Horowitz S, Gorovsky MA: **An unusual genetic code in nuclear genes of *Tetrahymena*.** *Proc Natl Acad Sci USA* 1985, **82:**2452-2455.
7. Hanyu N, Kuchino Y, Susumu N, Beier H: **Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* tRNAs^{Gln}.** *EMBO J* 1986, **5:**1307-1311.
8. Liang A, Heckmann K: ***Blepharisma* uses UAA as a termination codon.** *Naturwissenschaften* 1993, **80:**225-226.
9. Osawa S, Jukes TH, Watanabe K, Muto A: **Recent evidence for evolution of the genetic code.** *Microbiol Rev* 1992, **56:**229-264.
10. Knight RD, Freeland SJ, Landweber LF: **Rewiring the keyboard: evolvability of the genetic code.** *Nat Rev Genet* 2001, **2:**49-58.
11. Baroin Tourancheau A, Tsao N, Klobutcher LA, Pearlman RE, Adoutte A: **Genetic code deviations in the ciliates: evidence for multiple and independent events.** *EMBO J* 1995, **14:**3262-3267.
12. Preer JR Jr, Preer LB, Rudman BM, Barnett AJ: **Deviations from the universal code shown by the gene for surface protein 51A in** *Paramecium*. *Nature* 1985, **314:**188-190.
13. Meyer F, Schmidt HJ, Plumper E, Hasilik A, Mersmann G, Meyer HE, *et al.*: **UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*.** *Proc Natl Acad Sci USA* 1991, **88:**3758-3761.
14. Caron F: **Eucaryotic codes.** *Experientia* 1990, **46:**1106-1117.

15. Grimm M, Brunen-Nieweler C, Junker V, Heckmann K, Beier H: **The hypotrichous ciliate *Euplotes octocarinatus* has only one type of tRNA^{Cys} with GCA anticodon encoded on a single macronuclear DNA molecule.** *Nucleic Acids Res* 1998, **26:**4557-4565.
16. Dontsova M, Frolova L, Vassilieva J, Piendl W, Kisselev L, Garber M: **Translation termination factor aRF1 from the archaeon Methanococcus jannaschii is active with eukaryotic ribosomes.** *FEBS Lett* 2000, **472:**213-216.
17. Karamyshev AL, Ito K, Nakamura Y: **Polypeptide release factor eRF1 from Tetrahymena thermophila: cDNA cloning, purification and complex formation with yeast eRF3.** *FEBS Lett* 1999, **457:**483-488.
18. Song H, Mugnier P, Das AK, Webb HM, Evans DR, Tuite, MF, Hemmings BA, Barford D: **The crystal structure of human eukaryotic release factor eRF1 – Mechanism of stop codon recognition and peptidyl-tRNA hydrolysis.** *Cell* 2000, **100:**311-321.
19. Mironova LN, Zelenaia OA, Ter-Avanesian MD: **Nuclear-mitochondrial interactions in yeasts: mitochondrial mutations compensating the respiration deficiency of sup1 and sup2 mutants.** *Genetika* 1986, **22:**200-208.
20. Knight RD, Landweber LF: **The early evolution of the genetic code.** *Cell* 2000, **101:**569-572.
21. Lynn DH, Small EB: **A revised classification of the Phylum Ciliophora Doflein, 1901.** *Rev Soc Mex Hist Nat* 1997, **47:**65-78.
22. Baroin Tourancheau A, Villalobo E, Tsau N, Torres A, Pearlman RE: **Protein coding gene trees in ciliates: comparison with rRNA-based phylogenies.** *Mol Phylogenet Evol* 1998, **10:**299-309.
23. Stechmann A, Schlegel M, Lynn DH: **Phylogenetic relationships between Prostome and Colpodean ciliates tested by small subunit rRNA sequences.** *Mol Phylogenet Evol* 1998, **9:**48-54.
24. Wright AG, Dehority BA, Lynn DH: **Phylogeny of the rumen ciliates *Entodinium, Epidinium* and *Polyplastron* (Litostomatea: Entodiniomorphida) inferred from small subunit ribosomal RNA sequences.** *J Eukaryot Microbiol* 1997, **44:**61-67.
25. Strüder-Kypke MC, Wright AG, Fokin SI, Lynn D: **Phylogenetic relationships of the subclass Peniculia (Oligohymenophorea, Ciliophora) inferred from small subunit rRNA gene sequences.** *J Eukaryot Microbiol* 2000, **47:**419-429.
26. Small EB, Lynn DH: **A new macrosystem for the Phylum Ciliophora Doflein, 1901.** *Biosystems* 1981, **14:**387-401.
27. Corliss JO: *The Ciliated Protozoa. Characterization, Classification, and Guide to the Literature.* London: Pergamon Press; 1979.
28. Hammerschmidt B, Schlegel M, Lynn DH, Leipe DD, Sogin ML, Raikov IB: **Insights into the evolution of nuclear dualism in the ciliates revealed by phylogenetic analysis of rRNA sequences.** *J Eukaryot Microbiol* 1996, **43:**225-230.
29. Lynn DH, Corliss JO: **Ciliophora**. In *Microscopic Anatomy of Invertebrates. Vol. 1: Protozoa.* Edited by Corliss JO, Harrison FW. New York: Wiley-Liss, Inc.; 1991:333-467.
30. Bertram G, Bell HA, Ritchie DW, Fullerton G, Stansfield I: **Terminating eukaryotic translation: domain 1 of release factor eRF1 functions in stop codon recognition.** *RNA* 2000, **6:**1236-1247.
31. Liang A, Brünen-Nieweler C, Muramatsu T, Kuchino Y, Beier H, Heckmann K: **The ciliate *Euplotes octocarinatus* expresses two polypeptide release factors of the type eRF1**. *Gene* 2001, **262:**161-168.
32. Schneider SU, Leible MB, Yang XP: **Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of Acetabularia and the occurrence of unusual codon usage.** *Mol Gen Genet* 1989, **218:**445-452.
33. Lovett PS, Ambulos NP Jr, Mulbry W, Noguchi N, Rogers, EJ: **UGA can be decoded as tryptophan at low efficiency in *Bacillus subtilis*.** *J Bacteriol* 1991, **173:**1810-1812.
34. Matsugi J, Murao K, Ishikura H: **Effect of *B. subtilis* tRNA(Trp) on readthrough rate at an opal UGA codon.** *J Biochem* 1998, **123:**853-858.
35. Schultz DW, Yarus M: **Transfer RNA mutation and the malleability of the genetic code.** *J Mol Biol* 1994, **235:**1377-1380.
36. Woese CR, Fox GE, Zablen L, Uchida T, Bonen L, Pechman K, *et al.*: **Conservation of primary structure in 16S ribosomal RNA.** *Nature* 1975, **254:**83-86.
37. Curtis EA, Landweber LF: **Evolution of gene scrambling in ciliate micronuclear genes.** *Ann NY Acad Sci* 1999, **870:**349-350.
38. Horton TL, Landweber LF: **Evolution of four types of RNA editing in myxomycetes.** *RNA* 2000, **6:**1339-1346.
39. Harvey PH, Pagel MD: *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press; 1991.
40. Sokal RR, Rohlf FJ: *Biometry: The Principles and Practice of Statistics in Biological Research.* New York: W. H. Freeman and Company; 1995.
41. Inagaki Y, Doolittle WF: **Evolution of the eukaryotic translation termination system: origins of release factors.** *Mol Biol Evol* 2000, **17:**882-889.